Bioinformatic solutions for chromosomal copy number analysis in cancer

Scheinin, I.

2017

document version
Publisher's PDF, also known as Version of record

Abstract

Chromosomal copy number aberrations are one of the main mechanisms that give rise to the proliferative capabilities of cancer cells. These aberrations can be quantified with technologies that generate measurements genome-wide and with high resolution. Hence, they produce vast amounts of data, which requires tailored bioinformatic solutions for analysis and management. Two such high-resolution and genome-wide technologies are DNA microarrays, which are successively replaced by next-generation sequencing approaches. This dissertation describes three novel bioinformatic solutions for copy number analysis in cancer with these technologies.

CanGEM is a publicly-accessible database solution for storage of raw and processed copy number data from cancer research experiments. The contents of the database can be queried based on clinical and copy number data. Clinical data is collected using appropriate controlled vocabularies. Copy number data is collected as raw microarray data and automated analysis identifies the locations of chromosomal aberrations. In order to allow integration of data measured with different microarray platforms, a copy number status is derived for every known human gene.

CGHpower is a statistical power calculator for copy number experiments that compare two groups. It estimates genome complexity of a cancer type in question from a pilot data set of the sample series, and assesses the number of samples required to satisfy statistical requirements. It can be used either in the planning stages of experiments, including as a justification in grant applications, or to verify whether sufficient samples were included in past experiments. Performance of this bioinformatic solution is evaluated with real and simulated data sets.

QDNAseq is a preprocessing solution to detect copy number aberrations from shallow whole-genome next-generation sequencing data. It corrects the observed sequencing coverage for known systematic biases and allows filtering of spurious regions in the genome. A new list of such problematic regions is derived from public data generated by the 1000 Genomes Project. Performance of the solution is evaluated relative to other similar published solutions and DNA microarrays, and also compared to theoretical statistical expectations.

An application of the QDNAseq method is also presented in a translational research project with the aim to identify copy number aberrations in tumors of patients with low-grade glioma. Aberrations identified by shallow whole-genome next-generation sequencing and QDNAseq are used to evaluate associations with patient survival, and also to assess intratumoral heterogeneity and temporal evolution of these tumors. A loss in chromosome 10q is identified to be associated with poor prognosis, and the finding validated in two independent data sets. From the assessment of intratumoral heterogeneity and temporal tumor evolution, the well-characterized co-deletion of 1p/19q is found to be the only chromosomal aberration that is consistently present or absent across the entire tumor and possible future recurrences. This is compatible with the present view of its role as an early event in the development of these tumors.

The text concludes with a discussion of lessons learned from the development process and application of the three described bioinformatic solutions. Better awareness of and adherence to established best practices from the software development field would have been useful, and together with more careful consideration of implementation decisions could have resulted in software that was more apt for its purpose while also more efficient to develop and maintain. Similar to the presented solutions, much of the development of custom bioinformatics software is performed within academic research groups. Closer attention to the software development process itself could possibly be beneficial for academic software development in general.